ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

УДК 656.073, 004.934 doi:10.20998/2413-4295.2018.45.14

ПОРІВНЯННЯ ЕФЕКТИВНОСТІ ДВОХ МЕТОДІВ ФОРМАЛІЗАЦІЇ ГОЛОСОВОЇ ВЗАЄМОДІЇ

І. М. НАЙДЬОНОВ

кафедра технологій управління, Київський національний університет імені Тараса Шевченка, Київ, УКРАЇНА
e-mail: samogot@gmail.com

АНОТАЦІЯ *Стаття присвячена дослідженню ефективності формалізації голосової взаємодії без перетворення голосової інформації в текст, на основі застосування рефлекторної системи голосового управління, що складаються з фонемного стенографа, який перетворює звуковий запис на фонемну репрезентацію, і ядра класифікації, яке визначає зміст та керуючий вплив з отриманого набору фонем. Мета статті полягає у порівнянні ефективності методів машинного навчання для формалізації голосової взаємодії на прикладі системи підтримки диспетчеризації автотранспорту з використанням рефлекторної системи голосового управління. З метою перевірки ефективності побудованих моделей було проведено ітеративний процес збору даних (у відповідності до моделі голосової взаємодії у вигляді дерева сценаріїв) та моделювання формалізації, який передбачав аналіз отриманих результатів та розширення метрик точності оцінювання для незбалансованих вибірок (прецизійність, повнота, F-міра). На первинному етапі зібрано голосові дані 23 дикторів, у середньому по 45 зразків на реакцію. Результати моделювання на мінімальному наборі даних обома методами показали точність не вищу за 50%, що є недостатньою для практичного застосування. На основі цього було висунуто гіпотезу про малу кількість голосових даних для машинного навчання, тому на другому етапі зібрано в середньому 310 голосових зразків для кожної з 3-х реакцій простого контексту, загалом 925 реакцій. Моделювання методом інтелектуальних рефлекторних систем показало точність біля 60%, що також є недостатнім, а методом згорткових нейронних мереж — трохи більше за 90%, що є прийнятним. Для підтвердження ефективності методу інтелектуальних рефлекторних систем двох ітерацій виявилося недостатньо, висунуто гіпотезу про недостатню якість звукового запису та високий рівень шумів як перешкоди ефективності моделі формалізації, окреслено перспективи проведення наступного етапу дослідження. Зроблено висновок про ефективність рефлекторної системи голосового управління та її здатність на практиці визначати зміст та керуючий вплив отриманого набору фонем без перетворення голосової інформації в текстову форму.*
Ключові слова: *інтелектуальні рефлекторні системи; згорткові нейронні мережі; класифікація голосових команд; класифікація мовлення; розпізнавання мовлення; обробка природної мови*

COMPARISON OF THE EFFECTIVENESS OF TWO METHODS OF FORMALIZATION OF VOICE INTERACTION

I. NAYDONOV

Department of Technology Management, Taras Shevchenko National University of Kyiv, Kyiv, UKRAINE

ABSTRACT *The article is devoted to the study of the effectiveness of formalization of voice interaction without the transformation of voice information into text, based on the use of a reflex voice control system consisting of a phonemic transcript that converts a sound recording to a phonemic representation, and a classification core that determines the content and control of the received phonemic set. The purpose of the paper is to compare the effectiveness of machine learning methods for formalizing voice interaction on an example of a support system for vehicle dispatching using a reflex voice control system. In order to verify the effectiveness of the constructed models, an iterative process of data collection (in accordance with the model of voice interaction in the form of a tree of scenarios) and formalization modeling was carried out, which included analysis of the results and the expansion of metrics for the accuracy of the evaluation for unbalanced samples (precision, recall, F-score). At the initial stage, voice data of 23 speakers was collected, with an average of 45 samples per reaction. The simulation results on a minimum set of data by both methods showed an accuracy of no more than 50%, which is insufficient for practical application. On the second stage, an average of 310 voice samples were collected for each of the 3 simple-context reactions, a total of 925 reactions. The simulation by the method of intelligent reflex systems showed a accuracy of about 60%, which is also insufficient, and the accuracy of method of convolutional neural networks is slightly more than 90%, which is acceptable. In order to confirm the efficiency of the method of intelligent reflex systems, two stages was insufficient, the hypothesis about insufficient quality of sound recordings and high level of noise as obstacles to the effectiveness of the formalization model was advanced, prospects for conducting the next stage of the research were outlined. A conclusion is made about the effectiveness of the reflex voice control system and its ability to determine in practice the content and control of the received phonemic set without converting the voice information into a text form.*
Keywords*: intelligent reflex systems; convolutional neural networks; voice commands classification; speach classification; speach recognition; natural language processing*

Вступ

Голосова взаємодія – важливий аспект життя кожної людини. Автоматизація цієї голосової взаємодії інтенсивно розвивається, особливо в таких галузях як робототехніка [1], Інтернет речей (IOT) та через автоматизацію багатьох процесів, зокрема у транспорті [2,3] та дистрибуції [4,5]. Запровадження голосових інтерфейсів допомагає звільнити руки від

управління технічними пристроями та зробити цей процес більш зручним і природним для людини.

Формалізація голосової взаємодії є достатньо важливим напрямом. У більшості систем це досягається переведенням голосової інформації у текст з подальшим використанням. Такий підхід є достатньо розповсюдженим та існує велика кількість готових систем розпізнавання голосу у текст, але для забезпечення високої якості, такі системи потребують великої кількості ресурсів, тому зазвичай вони встановлюються на серверах, а для їх використання необхідний постійний доступ до мережі Інтернет для зв'язку з ними. Крім того, для використання систем формалізації голосової інформації в автоматизованих системах управління недостатньо перевести голосову інформацію в текст, необхідно ще забезпечити виділення керуючого впливу з отриманої команди.

Існують системи голосового управління, які не потребують розпізнавання голосової інформації і перетворення її в текстову форму, а можуть одразу визначити керуючий вплив з вимовленої команди. Так наприклад, рефлекторні системи голосового управління [6-8] на основі теорії несилової взаємодії [9] працюють за таким принципом і складаються з двох основних частин: фонемного стенографа [10], який перетворює звуковий запис на фонемну репрезентацію, і ядра класифікації, яке визначає зміст та керуючий вплив з отриманого набору фонем.

Для автоматизації голосової взаємодії можуть бути використані певні, заздалегідь створені, сценарії голосової взаємодії [11], які допоможуть системі визначити зміст сказаного. Створення подібних сценаріїв взаємодії має сенс для технологічних ситуацій, в яких взаємодія відбувається за певним протоколом, наприклад, у промислових системах, перемовинах з диспетчерами в різних галузях: медицина, авіація [12], правоохоронні органи чи дистрибуція [11]. Створення сценаріїв голосової взаємодії для вільного спілкування людей не має наразі настільки важливого економічного значення, але такі спроби теж існують, наприклад для подолання сором'язливості [1].

Центральна ланка формалізації голосової взаємодії – це класифікація вимовлених команд серед заздалегідь створених можливих сценаріїв взаємодії. Для виконання цієї класифікації можуть бути використані різні методи машинного навчання.

Мета роботи

Мета статті полягає у порівнянні ефективності методів машинного навчання для формалізації голосової взаємодії на прикладі системи підтримки диспетчеризації автотранспорту з використанням рефлекторної системи голосового управління.

Викладення основного матеріалу

Рефлекторні системи голосового управління складаються з двох основних частин: фонемного стенографа [10], який перетворює звуковий запис на фонемну репрезентацію, і ядра класифікації, яке визначає зміст та керуючий вплив кожного з отриманих наборів фонем. У якості ядра класифікації було запропоновано дуальну систему класифікації фонемної репрезентації голосових команд [13], яка може використовувати метод інтелектуальних рефлекторних систем (ІРС) [6] або метод згорткових нейронних мереж (ЗНМ) для різних предметних областей, в залежності від того, який показує кращі результати.

Згорткові нейронні мережі для роботи з фонемами найбільше нагадують використання методу згорткових нейронних мереж для задачі класифікації текстів [14], але оперують з «текстом» не за словами, а пофонемно, що схоже на роботу з текстом посимвольно [15].

Була розроблена модель голосової взаємодії водія та системи підтримки диспетчеризації автотранспорту у вигляді дерева сценаріїв [11], яка складається із 64 основних команд (реакцій) та поділяє всю множину голосової взаємодії на 19 контекстів, що можуть бути змодельовані окремо.

З метою перевірки ефективності побудованих моделей було проведено ітеративний процес збору даних та моделювання, який передбачав аналіз отриманих результатів та введення нових критеріїв оцінки, якщо попередні не давали достатньої точності оцінювання. У цілому цей процес апробації засобів формалізації голосової інформації в системах диспетчерського контролю за рухом автотранспорту можна розбити на два основні етапи.

На першому етапі була зібрана широка вибірка голосових даних згідно із моделлю дерева сценаріїв голосової взаємодії в системах підтримки диспетчеризації автотранспорту. Зібрані голосові дані можна охарактеризувати наступними параметрами:

- 4 пристрої, 23 диктори (11 жінок, 12 чоловіків), 94 варіанти стимулів (64 на основні реакції), 3046 зразків;
- додатково 23 варіанти стимулів та 465 голосових зразків для реакції розпізнавання часу.

Детальний розподіл голосових даних, зібраних на першому етапі моделювання, представлено в таблиці 1. Ця таблиця показує розподіл зібраних голосових зразків за пристроями та дикторами, а також повідомляє загальну кількість голосових зразків надиктованих кожним диктором, кількість унікальних реакцій та варіантів формулювання стимулів різними словами серед зібраних голосових зразків. В окрему колонку виведено кількість додаткових записів для реакції розпізнавання часу.

Як можна бачити, не всі диктори надиктували повний перелік реакцій згідно до моделі голосової взаємодії в системах підтримки диспетчеризації. Це має сенс, оскільки не всі водії будуть стикатися з усіма можливими позаплановими подіями. Також деякі диктори надиктували повністю ідентичні фрази

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

кілька разів, що б мати можливість виключити вплив сторонніх шумів і певної стохастичності вимови та запису звуку, а деякі диктори надиктували різні варіанти формулювання та вимови стимулів, що б мати можливість навчити систему виділяти ключову інформацію в різних формулюваннях природної мови.

Таблиця 1 – Детальний розподіл за дикторами та пристроями голосових даних зібраних на першому етапі моделювання

Диктор	Пристрій	Стать	Кількість записів	Кількість реакцій	Кількість варіантів стимулів	Кількість записів реакції часу
1	1	жін.	109	64	92	22
2	1	жін.	105	64	96	22
3	1	жін.	100	64	96	21
4	1	жін.	97	64	93	22
5	1	жін.	96	64	93	22
6	1	жін.	95	64	94	22
7	1	жін.	95	64	95	23
8	1	жін.	95	62	90	21
9	1	жін.	90	62	89	22
10	1	жін.	79	58	79	23
11	1	чол.	196	64	96	44
12	1	чол.	101	64	94	25
13	1	чол.	96	64	92	22
14	1	чол.	98	63	95	21
15	1	чол.	89	63	85	22
16	1	чол.	98	62	90	21
17	1	чол.	64	48	62	22
18	1	чол.	30	25	30	0
19	1	чол.	23	16	22	0
20	1	чол.	23	9	19	0
21	2	жін.	97	64	94	22
22	3	чол.	96	64	96	22
23	4	чол.	99	64	96	24

Першим було проведене моделювання методом інтелектуальних рефлекторних систем з розміром N-грам 1–3. Результати моделювання наведено в таблиці 2. Було створено окрему модель для кожного контексту взаємодії водія та системи підтримки диспетчеризації автотранспорту, а також модель розпізнавання реакцій для всієї вибірки, без використання контекстів. Крім контекстів, представлених у моделі голосової взаємодії, було використано додатковий тестовий контекст, який складався з трьох реакцій найпростішого дерева сценаріїв голосової взаємодії [11].

Розміри N-грам 1–3 було обрано, виходячи з того, що метод інтелектуальних рефлекторних систем має квадратичну залежність швидкості розрахунку від максимального розміру N-грам. З таблиці 2 видно, що точність розпізнавання цих моделей не перевищує 50% для жодного з контекстів, крім першого. Тому для підвищення якості моделей було проведено моделювання цим же методом, але з розміром N-грам 2–4. Результати цього моделювання представлено в таблиці 3.

Таблиця 2 – Результати моделювання першого набору даних використовуючи IPC з послідовностями розміром 1–3

№ Контексту	Точність розпізнання	Середня прецизійність	Середня повнота	Середня F-міра	Кількість записів
1	0.677	0.450	0.445	0.447	124
2	0.456	0.559	0.663	0.408	513
3	0.387	0.055	0.143	0.080	217
4	0.375	0.217	0.271	0.171	104
5	0.339	0.108	0.184	0.111	183
6	0.397	0.198	0.252	0.222	209
7	0.206	0.021	0.098	0.034	233
8	0.346	0.182	0.179	0.107	217
9	0.344	0.144	0.225	0.141	131
10	0.349	0.489	0.255	0.194	126
11	0.268	0.099	0.145	0.114	239
12	0.264	0.303	0.250	0.238	201
13	0.450	0.361	0.259	0.182	171
14	0.439	0.551	0.432	0.432	139
15	0.302	0.547	0.210	0.152	159
16	0.478	0.539	0.470	0.450	115
17	0.271	0.261	0.263	0.248	155
18	0.418	0.482	0.411	0.407	110
19	0.471	0.650	0.458	0.462	87
По всій вибірці	0.044	0.027	0.017	0.009	2069
Тестовий контекст	0.485	0.162	0.333	0.218	101

Як альтернативний класифікатор у дуальній системі класифікації фонемної репрезентації голосових команд використано метод згорткових нейронних мереж [13]. На відміну від методу інтелектуальних рефлекторних систем, метод згорткових нейронних мереж залежить від розміру фільтрів лише лінійно, оскільки замість всіх можливих наборів N-грам, використовується лише певна фіксована кількість фільтрів, які навчаються методом зворотного розповсюдження помилки. Тому для моделювання методом згорткових нейронних мереж використовувалися лише фільтри розміром 2–4. Результати цього моделювання представлено в таблиці 4.

Таблиці 2, 3 та 4 включають деякі міри оцінки якості моделей та кількість звукових записів, які використовувалися для навчання та оцінки моделей.

Спочатку в якості основної міри використовувалася точність (accuracy) [16], що може бути розрахована за наступною формулою:

$$A = \frac{\sum_i TP_i}{\sum_i TP_i + FN_i}, \quad (1)$$

де A — точність, TP_i — кількість вірно розпізнаних (True positive) зразків реакці i, а FN_i — кількість зразків реакції i, хибно розпізнаних (False negative) як інша реакція.

Таблиця 3 – Результати моделювання першого набору даних використовуючи IPC з послідовностями розміром 2–4

№ Контексту	Точність розпізнання	Середня прецизійність	Середня повнота	Середня F-міра	Кількість записів
1	0.710	0.503	0.442	0.459	124
2	0.895	0.465	0.478	0.471	513
3	0.382	0.048	0.124	0.069	217
4	0.558	0.740	0.480	0.484	104
5	0.415	0.434	0.259	0.225	183
6	0.565	0.282	0.359	0.316	209
7	0.232	0.388	0.127	0.089	233
8	0.373	0.410	0.207	0.157	217
9	0.527	0.737	0.437	0.449	131
10	0.532	0.773	0.468	0.486	126
11	0.515	0.207	0.282	0.238	239
12	0.488	0.510	0.469	0.452	201
13	0.485	0.448	0.303	0.283	171
14	0.633	0.703	0.625	0.625	139
15	0.453	0.480	0.329	0.333	159
16	0.583	0.661	0.574	0.550	115
17	0.387	0.473	0.378	0.374	155
18	0.582	0.672	0.576	0.583	110
19	0.655	0.728	0.650	0.642	87
По всій вибірці	0.122	0.061	0.048	0.027	2069
Тестовий контекст	0.475	0.160	0.327	0.215	101

Оскільки частота реакцій у більшості контекстів не збалансована, точність може показувати кращі результати, ніж справжня якість моделі.

Побудувавши графіки розподілу точності та F-міри від кількості зразків у контексті (рис. 1) ми можемо бачити, що прямої залежності немає. Але на якість розпізнавання повинна впливати кількість реакцій в контексті і середня кількість записів зразків голосових даних на кожну реакцію.

З матриці помилок (confusion matrix) [17] моделювання тестового контексту методом інтелектуальних рефлекторних систем (рис. 2а та 2б), ми можемо бачити, що рівень розпізнавання близький до 50% досягається за рахунок того, що всі реакції в контексті розпізнаються, як реакція №2. Оскільки вона представлена найбільшою кількістю зразків, точність досягає відносно високих значень при насправді низькій якості моделі.

Таблиця 4 – Результати моделювання першого набору даних використовуючи ЗНМ з послідовностями розміром 2–4

№ Контексту	Точність розпізнання	Середня прецизійність	Середня повнота	Середня F-міра	Кількість записів
1	0.879	0.880	0.863	0.870	124
2	0.957	0.933	0.799	0.851	513
3	0.820	0.870	0.777	0.813	217
4	0.846	0.857	0.829	0.837	104
5	0.798	0.814	0.734	0.754	183
6	0.852	0.851	0.789	0.814	209
7	0.785	0.808	0.773	0.785	233
8	0.871	0.895	0.847	0.867	217
9	0.863	0.871	0.836	0.847	131
10	0.825	0.841	0.818	0.819	126
11	0.908	0.922	0.850	0.878	239
12	0.846	0.852	0.849	0.847	201
13	0.848	0.860	0.842	0.849	171
14	0.842	0.851	0.838	0.838	139
15	0.849	0.842	0.829	0.834	159
16	0.817	0.826	0.814	0.814	115
17	0.723	0.729	0.723	0.720	155
18	0.836	0.842	0.837	0.835	110
19	0.862	0.871	0.863	0.856	87
По всій вибірці	0.652	0.679	0.622	0.640	2069
Тестовий контекст	0.891	0.915	0.871	0.888	101

Дослідивши матриці помилок моделювання інших контексів методом інтелектуальних рефлекторних систем було виявлено, що моделі всіх контекстів в тій чи іншій мірі схильні до подібних помилок. Найменш схильними виявилися моделі контекстів 14, 16, 18 та 19, але за таблицею ми можемо бачити, що значення точності для цих контекстів менші за точність моделей для 13-го та тестового контекстів, де всі реакції були розпізнані як одна й та ж сама (рис. 2а).

Як кращі міри якості класифікаційної моделі для незбалансованих вибірок зазвичай використовують прецизійність (precision) та повноту (recall), які у випадку небінарної класифікації можуть бути визначені лише для певного класу, а не для моделі в цілому [18]. Для оцінки якості моделі було використано середні показники прецизійності та

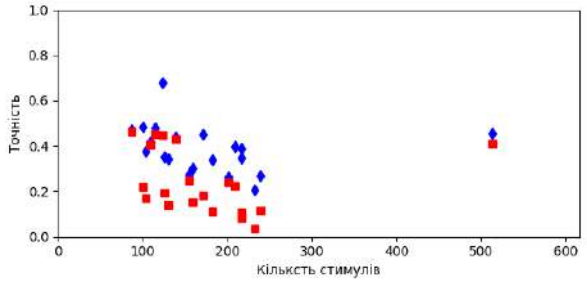
ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

повноти по всім реакціям, що можуть бути визначені по формулам:

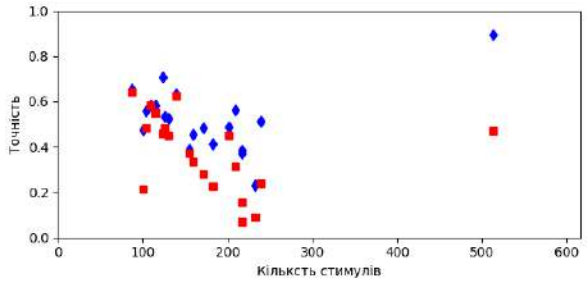
$$P = \frac{1}{n}\sum_i \frac{TP_i}{TP_i+FP_i}; \quad (2)$$

$$R = \frac{1}{n}\sum_i \frac{TP_i}{TP_i+FN_i}, \quad (3)$$

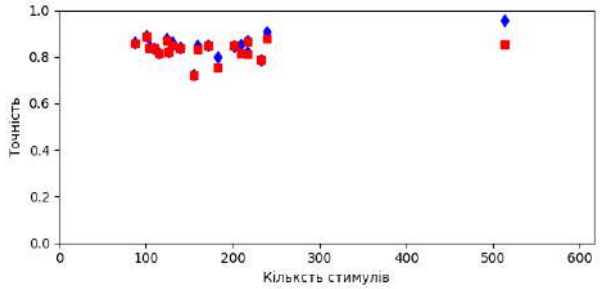
де P — середня прецизійність, R — середня повнота, n — кількість класів TP_i — кількість вірно розпізнаних (True positive) зразків реакції i, FP_i — кількість зразків хибно розпізнаних (False positive) як реакція i, а FN_i – кількість зразків реакції i, хибно розпізнаних (False negative) як інша реакція.



(а) Метод ІРС розміром 1–3

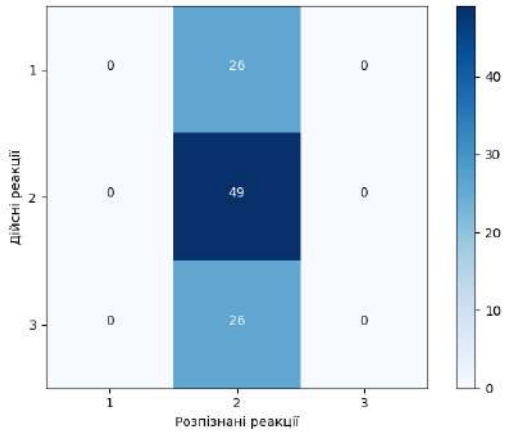


(б) Метод ІРС розміром 2–4

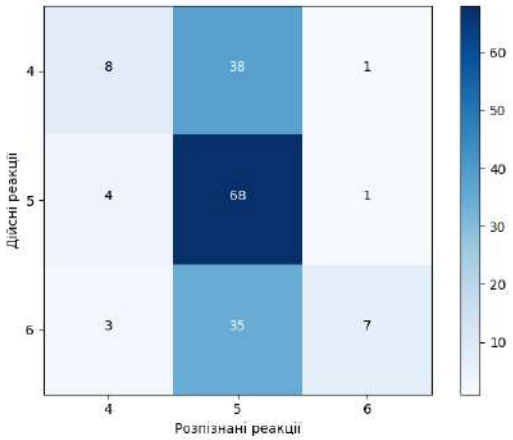


(в) Метод ЗНМ

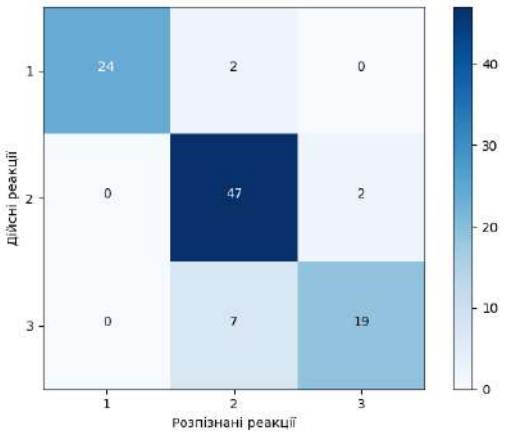
Рис. 1 – Розподіл точності (червоні квадрати) та F-міри (сині ромби) за кількістю голосових зразків при моделюванні контекстів різними методами



(а) Метод ІРС розміром 1–3



(б) Метод ІРС розміром 2–4



(в) Метод ЗНМ

Рис. 2 – Порівняння матриць помилок трьох різних методів (а, б, в) розпізнавання по реакціях для тестового контексту першого набору даних

Оскільки прецизійність показує лише помилки першого роду, а повнота лише помилки другого роду, існує узагальнена F-міра (F1, F-score) [18,19], що враховує обидва типи помилок як середньо-гармонічне, та може бути визначена за формулою:

$$F_1 = 2\frac{P \cdot R}{P+R}, \quad (1)$$

де F_1 — F-міра, P — прецизійність а R — повнота.

Дослідивши значення F-міри з таблиці 2 ми бачимо, що саме моделі контекстів 14, 16, 18 та 19, які виглядають найкраще на матрицях помилок, мають одні з найбільших значень F-міри, та порівняні зі значеннями моделей контекстів 1 та 2, що складаються лише з двох реакцій. Моделювання методом IPC з розміром N-грам 2–4 (табл. 3) показало схожі результати і дало невеликий приріст. якості розпізнавання, але цієї якості все одно недостатньо для практичного застосування моделі.

Метод згорткових нейронних мереж показав кращий результат. З таблиці 4 видно, що значення F-міри не набагато нижчі за точність, з чого можна зробити висновок, що ця модель краще працює з незбалансованими вибірками.

Для розрахунку показників у таблицях 2–4 використовувався метод кросс-валідації [20] з розбиттям повної вибірки на 5 рівних випадкових частин. Таким чином моделювання проводилося 5 разів, так, щоб кожна з 5 частин один раз була використана як тестова вибірка, а 4 інших частини в кожному моделюванні складали навчальну вибірку. Саме комбінація результатів на тестовій вибірці з 5-ти моделювань і була використана для розрахунку метрик в таблицях.

Розрахунок метрик на тестових вибірках показував точність розпізнавання від 90% до 100%, це свідчить про ефект перенавчання системи. Одним зі способів вирішення проблеми перенавчання є збільшення кількості даних. Було висунуто гіпотезу про низьку якість розпізнавання в зв'язку з недостатньою кількістю вхідних даних.

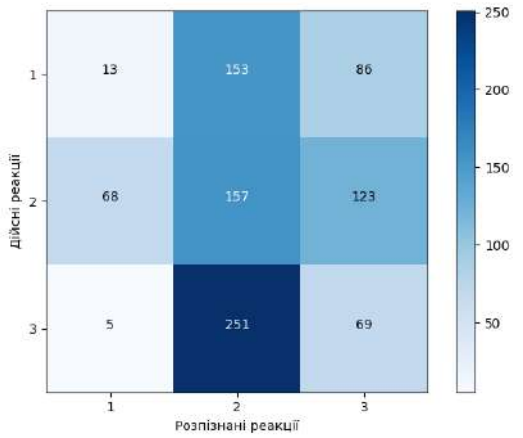
Отже, було вирішено провести другий етап дослідження, для якого було необхідно зібрати більшу кількість голосових даних. Для перевірки гіпотези вирішено зібрати голосові зразки лише для одного тестового контексту.

Додаткові голосові дані зібрано для одного контексту: 1 пристрій, 1 диктор (чоловік), 37 варіантів стимулів, 3 реакції у контексті, тобто 938 зразків.

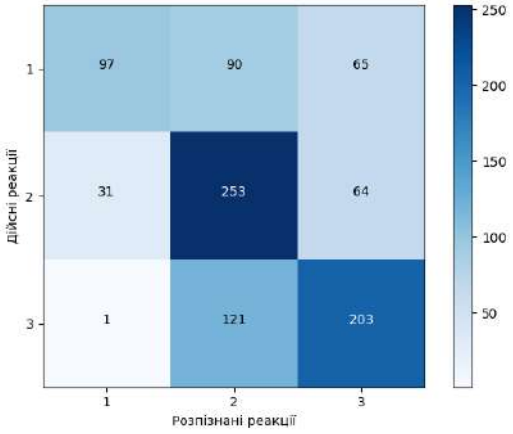
Результати моделювання другого набору даних трьома різними методами ми можемо бачити в таблиці 5, а матриці помилок зображено на рис. 3.

Таблиця 5 – Порівняння якості розпізнавання другого набору даних різними методами

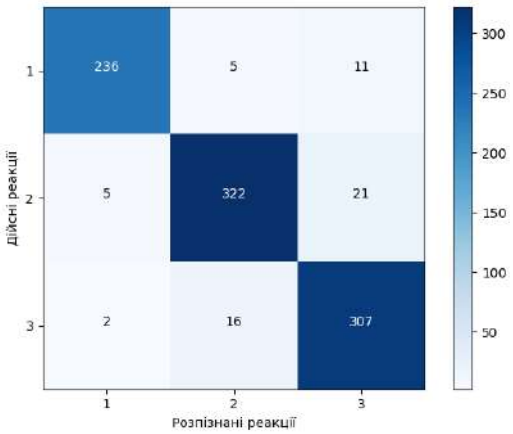
Показник	IPC 1–3	IPC 2–4	ЗНМ
Точність розпізнавання	0.258	0.598	0.935
Середня прецизійність	0.226	0.636	0.939
Середня повнота	0.238	0.579	0.935
Середня F-міра	0.217	0.583	0.937
Кількість зразків	925	925	925



(а) Метод IPC розміром 1–3



(б) Метод IPC розміром 2–4



(в) Метод ЗНМ

Рис. 3 – Порівняння матриць помилок трьох різних методів (а, б, в) розпізнавання по реакціях для тестового контексту другого набору даних

З результатів моделювання можна бачити, що збільшення кількості вхідних даних безумовно покращило якість розпізнавання, для всіх методів. Моделювання інтелектуальними рефлекторними

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

системами для одного контексту з великою вибіркою даних дало майже 60% точності та значення F-міри 0.58. З матриці помилок (рис. 3б) не має явного переважання певної реакції над іншими. На жаль, отриманих значень точності досі недостатньо для успішного використання моделі на практиці.

Моделювання методом згорткових нейронних мереж дало трохи більше 90% точності та відповідне значення F-міри. Така точність є достатньою для практичного застосування, отже можна зробити висновок, що за даних умов дуальна система класифікації фонемної репрезентації голосових команд працює краще з використанням методу згорткових нейронних мереж для побудови моделі формалізації голосової інформації в системах диспетчеризації автотранспорту.

Обговорення результатів

Порівняння різних метрик ефективності моделей небінарної класифікації показало, що для моделей високої якості, а також у випадках збалансованих вибірок даних, всі метрики показують схожі значення ефективності. Але у випадках моделей, які негативно реагують на незбалансованість вибірки, F-міра оцінює модель набагато точніше.

Було проведено 2 етапи моделювання. На первинному етапі зібрано голосові дані 23 дикторів, у середньому по 45 зразків на реакцію. Результати моделювання обома методами показали точність не вищу за 50%, що є недостатньою для практичного застосування. Точність класифікації на навчальних даних була близькою до 100%, що свідчить про перенавчання.

На основі цього було висунуто гіпотезу про недостатню кількість голосових даних, тому на другому етапі зібрано в середньому 310 голосових зразків для кожної з 3-х реакцій простого контексту. Моделювання методом інтелектуальних рефлекторних систем показало точність біля 60%, що також є недостатнім, а методом згорткових нейронних мереж – трохи більше за 90%, що є прийнятним.

З цього можна бачити, що за даних умов, ефективність методу згорткових нейронних мереж вища за метод інтелектуальних рефлекторних систем більш ніж на 30%. Ці результати викликають певний подив, оскільки в попередніх дослідженнях [5-7] метод інтелектуальних рефлекторних систем давав набагато кращі результати. Візуально дослідивши фонемну репрезентацію даних, ми помітили сильну зашумленість фонемних даних, хоча в звукових файлах рівень шуму був нижчим. Одне з можливих пояснень цього полягає в тому, що фонемний стенограф не був налаштований на використання мікрофону в мобільному телефоні, частотні показники якого можуть вносити перешкоди в роботу системи.

Тобто, висунуто наступну гіпотезу про недостатню якість звукового запису та високий рівень шумів як перешкоди ефективності моделі формалізації. Перспективою подальшого розвитку є проведення наступного етапу дослідження, для якого необхідно зібрати нову вибірку даних з використанням більш якісного зовнішнього мікрофона. Крім того, для усунення впливу незбалансованої вибірки даних, пропонується зібрати рівну кількість записів для кожної реакції з дерева сценаріїв голосової взаємодії.

Висновки

Перевірка моделювання розпізнавання команд на основі ітеративного процесу збору даних та введення нових критеріїв оцінки, якщо попередні не дали достатньої точності оцінювання, може забезпечити процес порівняння ефективності різних методів класифікації в дуальній моделі формалізації голосової взаємодії.

Доцільно використовувати набір метрик оцінки ефективності моделей класифікації, що обов'язково має включати крім оцінки точності ще робастні метрики для незбалансованої вибірки (такі, як прецизійність, повноту, F-міру) або візуальний аналіз матриць помилок.

Досягнуто прийнятний для практичного використання рівень точності в моделі, побудованої методом згорткових нейронних мереж при другій ітерації моделювання з достатньою кількістю навчальних даних.

Для підтвердження ефективності методу інтелектуальних рефлекторних систем двох ітерацій виявилося недостатньо, висунуто гіпотезу про недостатню якість звукового запису та високий рівень шумів як перешкоди ефективності моделі формалізації, окреслено перспективи проведення наступного етапу дослідження.

Загалом за результатами проведеної роботи підтверджено ефективність рефлекторної системи голосового управління, яка складається з фонемного стенографа і ядра класифікації, і здатна на практиці визначати зміст та керуючий вплив отриманого набору фонем без перетворення голосової інформації в текстову форму.

Список літератури

1. **Ishiguro, H.** Adaptation to teleoperated robots / **H. Ishiguro** // *International Journal of Psychology*. *31st International Congress of Psychology, 24–29 July 2016, Yokohama, Japan.* – 2016. – Vol. 51, issue S1. – P. 10. – doi:10.1002/ijop.12361.
2. **Кравченко, А. П.** Автоматизированная компьютерная система голосового управления автомобилем / **А. П. Кравченко, Н. М. Крамарь, И. В. Морозов** // *Автомобильный транспорт*. – 2009. – № 25. – С. 44-47.
3. **Heisterkamp, P.** Linguatronic Product-level Speech System for Mercedes-Benz Cars / **P. Heisterkamp** // *Proceedings of the First International Conference on Human Language*

Technology Research. – San Diego : Association for Computational Linguistics. – 2001. – P. 1-2. – doi:10.3115/1072133.1072199.

4. **Найдьонов, І. М.** Проблема голосової взаємодії в задачах управління дистрибуцією / **І. М. Найдьонов** // *Вісник Черкаського державного технологічного університету. Серія: Технічні науки*. – 2016. – № 3. – С. 63-71.
5. **Naydonov, I.** Geoinformation system of vehicle routing and parameters of voice interaction of subjects of logistics / **I. Naydonov** // *16th EAGE International Conference on Geoinformatics – Theoretical and Applied Aspects*. – 2017. – doi:10.3997/2214-4609.201701807.
6. **Егорченков, А. В.** Прикладное применение рефлекторной системы голосового управления / **А. В. Егорченков** // *Управління розвитком складних систем*. – 2016. – № 25. – С. 103-107.
7. **Teslia, I.** The Non-Force Interaction Theory for Reflex System Creation with Application to TV Voice Control / **I. Teslia, N. Popovych, V. Pylypenko, O. Chornyi** // *Proceedings of the 6th International Conference on Agents and Artificial Intelligence*. – 2014. – P. 288-296. – doi:10.5220/0004754702880296.
8. **Тесля, Ю. М**. Рефлекторная система голосового управления техническими устройствами (РСГУ) / **Ю. М. Тесля, О. Чорний** // *Управління розвитком складних систем*. – 2013. – № 15. – С. 105-110.
9. **Тесля, Ю. М.** Введення в інформатику природи / **Ю. М. Тесля**. – К. : Маклаут, 2010. – 255 с.
10. **Пилипенко, В. В.** Автоматизированный стенограф украинской речи / **В. В. Пилипенко, В. В. Робейко** // *Штучний інтелект*. – 2008. – № 4. – С. 768-775.
11. **Найдьонов, І. М.** Модель голосової взаємодії водія в системах диспетчерського контролю за рухом автотранспорту / **І. М. Найдьонов** // *Комп'ютерно інтегровані технології: освіта, наука, виробництво*. – 2018. – № 33. – С.127-127.
12. **Корсун, О. Н.** Экспериментальное исследование влияния акустических помех разных видов на результаты автоматического распознавания речевых команд / **О. Н. Корсун, А. А. Яцко, И. М. Финаев, В. Я. Чучупал** // *Наука и образование: научное издание МГТУ им. Н.Э. Баумана*. –2013. – Т. 1. – С. 12.
13. **Найдьонов, І. М.** Формалізація голосової інформації в системах диспетчерського контролю за рухом автотранспорту / **І. М. Найдьонов** // *Наукові нотатки*. – 2018. – № 64. – в друці.
14. **Kim, Y.** Convolutional Neural Networks for Sentence Classification / **Y. Kim** // *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. – 2014. – P. 1746-1751.
15. **Zhang, X.** Character-level Convolutional Networks for Text Classification / **X. Zhang, J. J. Zhao, Y. LeCun** // *Advances in Neural Information Processing Systems 28.* — 2015. — arXiv: 1509.01626.
16. **Ting, K. M.** Encyclopedia of machine learning / **K. M. Ting**. – Boston, MA : Springer, 2011. – 892 p.
17. **Stehman, S. V.** Selecting and interpreting measures of thematic classification accuracy / **S. V. Stehman** // *Remote Sensing of Environment*. — 1997. — Vol. 62, issue 1. — P. 77–89. — doi:10.1016/S0034-4257(97)00083-7.
18. **Powers, D. M. W.** Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation / **D. M. W. Powers** // *Journal of Machine Learning Technologies*. – 2011. – Vol. 2. – Issue 1. – P. 37-63.
19. **Sasaki, Y.** The truth of the F-measure / **Y. Sasaki**. – Manchester: University of Manchester, 2007. – 5 p. – doi:10.1007/978-0-387-30164-8.
20. **Kohavi, R.** A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection / **R. Kohavi** // *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann, 1995. – P. 1137-1143. – doi:10.1.1.48.529.

References (transliterated)

1. **Ishiguro, H.** Adaptation to teleoperated robots. *International Journal of Psychology,* 2016, **51**(S1), 10, doi:10.1002/ijop.12361.
2. **Kravchenko, A. P., Kramar, N. M., & Morozov, I. V.** Avtomatizirovannaja komp juternaja sistema golosovogo upravlenija avtomobilem [Automated computerized voice control system]. *Avtomobilnyj transport [Automobile transport]*, 2009, **25**, 44–47.
3. **Heisterkamp, P.** Linguatronic Product-level Speech System for Mercedes-Benz Cars. *Proceedings of the First International Conference on Human Language Technology Research*. San Diego: Association for Computational Linguistics, 2001, 1-2, doi:10.3115/1072133.1072199.
4. **Naydonov, I. M.** Problema holosovoi vzaiemodii v zadachakh upravlinnia dystrybutsiieiu [The problem of voice interaction in the distribution management tasks]. *Visnyk Cherkas koho derzhavnoho tekhnolohichnoho universytetu. Seriia: Tekhnichni nauky [Cherkasy state technological university journal. Series: Engineering],* 2016, **3**, 63–71.
5. **Naydonov, I.** Geoinformation system of vehicle routing and parameters of voice interaction of subjects of logistics. *16th eage international conference on geoinformatics - theoretical and applied aspects*, 2017, doi:10.3997/2214-4609.201701807.
6. **Egorchenkov, A. V.**. Prikladnoe primenenie reflektornoj sistemy golosovogo upravlenija [Applied application of the reflex voice control system]. *Upravlinnja rozvitkom skladnih sistem [Managing the development of complex systems]*, 2016, **25**, 103–107.
7. **Teslia, I., Popovych, N., Pylypenko, V., & Chornyi, O.** The non-force interaction theory for reflex system creation with application to tv voice control. *Proceedings of the 6th international conference on agents and artificial intelligence*, 2014, 288–296, doi: 10.5220/0004754702880296.
8. **Teslja, Ju. M., Chornij, O.** Reflektornaja sistema golosovogo upravlenija tehnicheskimi ustrojstvami (RSGU) [Reflex voice control system for technical devices (RVCS)]. *Upravlinnja rozvitkom skladnih sistem [Managing the development of complex systems]*, 2013, **15**, 105-110.
9. **Teslia, Yu. M.** Vvedennia v informatyku pryrody [Introduction to the informatics of nature]. K.: Maklaut, 2010, 255.
10. **Pilipenko, V. V., Robejko, V. V.** Avtomatizirovannyj stenograf ukrainskoj rechi [Automated stenographer of Ukrainian speech]. *Shtuchnij intelekt [Artificial Intelligence]*, 2008, **4**, 768–775.
11. **Naydonov, I. M.** Model holosovoi vzaiemodii vodiia v systemakh dyspetchers koho kontroliu za rukhom avtotransportu [Model of voice interaction of the driver in systems of dispatch control of motor transport]. *Komp iuterno-intehrovani tekhnolohii: osvita, nauka, vyrobnytstvo [Computer-integrated technologies: education, science, production]*, 2018, **33**, 121-127.

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

12. **Korsun, O. N., Jacko, A. A., Finaev, I. M., Chuchupal, V. Ja.** Eksperimental noe issledovanie vlijanija akusticheskih pomeh raznyh vidov na rezul taty avtomaticheskogo raspoznavanija rechevyh komand [Experimental study of the influence of acoustic noise of different types on the results of automatic recognition of voice commands]. *Nauka i obrazovanie: nauchnoe izdanie MGTU im. N.Je. Baumana [Science and education: a scientific publication MSTU. N.E. Bauman],* 2013, **1**, 12.
13. **Naydonov, I. M.** Formalizatsiia holosovoi informatsii v systemakh dyspetcherskoho kontroliu za rukhom avtotransportu [Formalization of voice information in systems of dispatch control over motor transport]. *Naukovi notatky [Scientific notes]*, 2018, **64** (in press).
14. **Kim, Y.** Convolutional neural networks for sentence classification. *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014, 1746–1751.
15. **Zhang, X., Zhao, J. J., LeCun, Y.** Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems 28*, 2015, arXiv: 1509.01626.
16. **Ting, K. M.** Encyclopedia of machine learning. Boston, MA: Springer, 2011, 892.
17. **Stehman, S. V.** Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment,* 1997, **62**(1), 77–89, doi:10.1016/S0034-4257(97)00083-7.
18. **Powers, D. M. W.** Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies,* 2011, **2**(1), 37-63.
19. **Sasaki, Y.** The truth of the f-measure. Manchester: University of Manchester, 2007, 5, doi:10.1007/978-0-387-30164-8.
20. **Kohavi, R.** A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the fourteenth international joint conference on artificial intelligence*. San Mateo, CA: Morgan Kaufmann, 1995, 1137–1143, doi:10.1.1.48.529.

Сведения об авторах (About authors)

Найдьонов Іван Михайлович – Київський національний університет імені Тараса Шевченка, аспірант кафедри технологій управління; м. Київ, Україна; ORCID: 0000-0002-2498-6375; e-mail: samogot@gmail.com.

Ivan Naydonov – PhD student, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine; ORCID: 0000-0002-2498-6375; e-mail: samogot@gmail.com.

Будь ласка, посилайтесь на цю статтю наступним чином:

Найдьонов, І. М. Порівняння ефективності двох методів формалізації голосової взаємодії / **І. М. Найдьонов** // *Вісник НТУ «ХПІ», Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». – 2018. – № 45 (1321). – С. 104-112. – doi:10.20998/2413-4295.2018.45.14.

Please cite this article as:

Naydonov, I. Comparison of the effectiveness of two methods of formalization of voice interaction. *Bulletin of NTU "KhPI". Series: New solutions in modern technologies*. – Kharkiv: NTU"KhPI", 2018, **45**(1321), 104–112, doi:10.20998/2413-4295.2018.45.14.

Пожалуйста, ссылайтесь на эту статью следующим образом:

Найдёнов, И. М. Сравнение эффективности двух методов формализации голосовой взаимодействия / **І. М. Найдёнов** // *Вестник НТУ «ХПИ», Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». – 2018. – № 45 (1321). – С. 104-112. – doi:10.20998/2413-4295.2018.45.14.

АННОТАЦИЯ *Статья посвящена исследованию эффективности формализации голосового взаимодействия без преобразования голосовой информации в текст, на основе применения рефлекторной системы голосового управления, состоящей из фонемного стенографа, который превращает звуковую запись в фонемную репрезентацию, и ядра классификации, которое определяет содержание и управляющее воздействие из полученного набора фонем. Цель статьи состоит в сравнении эффективности методов машинного обучения для формализации голосовой взаимодействия на примере системы поддержки диспетчеризации автотранспорта с использованием рефлекторной системы голосового управления. С целью проверки эффективности построенных моделей был проведен итеративный процесс сбора данных (в соответствии с моделью голосовой взаимодействия в виде дерева сценариев) и моделирования формализации, который предусматривал анализ полученных результатов и расширение метрик точности оценки для несбалансированных выборок (прецизионность, полнота F-мера). На первоначальном этапе собрано голосовые данные 23 дикторов, в среднем по 45 образцов на реакцию. Результаты моделирования на минимальном наборе данных обоими методами показали точность не выше 50%, что недостаточно для практического применения. На основе этого была выдвинута гипотеза о малом количестве голосовых данных для машинного обучения, поэтому на втором этапе собрано в среднем 310 голосовых образцов для каждой из 3-х реакций простого контекста, в общем 925 реакций. Моделирование методом интеллектуальных рефлекторных систем показало точность около 60%, что также является недостаточным, а методом сверточных нейронных сетей — чуть более 90%, что является приемлемым. Для подтверждения эффективности метода интеллектуальных рефлекторных систем двух итераций оказалось недостаточно, выдвинута гипотеза о недостаточном качестве звуковой записи и высокий уровень шумов как препятствия эффективности модели формализации, намечены перспективы проведения следующего этапа исследования. Сделан вывод об эффективности рефлекторной системы голосового управления и ее способность на практике определять содержание и управляющее воздействие полученного набора фонем без преобразования голосовой информации в текстовую форму.*

Ключевые слова: *интеллектуальные рефлекторные системы, сверточные нейронные сети, классификация голосовых команд; классификация речи; распознавание речи; обработка естественного языка*

Поступила (received) 23.11.2018